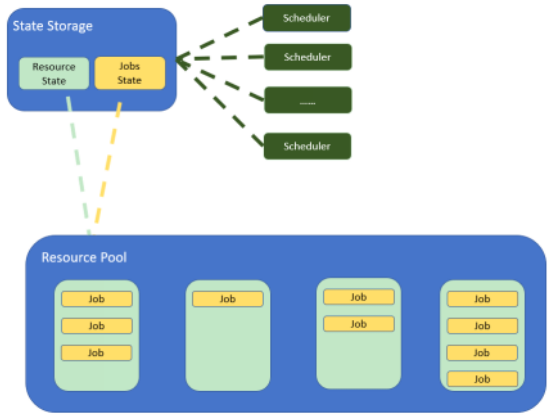
2.5 分布式调度架构之共享状态调度：物质文明、精神文明多手协商抓

2022年3月2日 10:35

两层调度问题.

第二层调度器只能看到部分资源. → { ① 不能保证全局状态的一致性. ② 不容易实现全局最优调度.

共享状态调度.



可以看出，**共享状态调度架构为了提供高可用性和可扩展性，将集群状态之外的功能抽出来作为独立的服务。**具体来说就是：

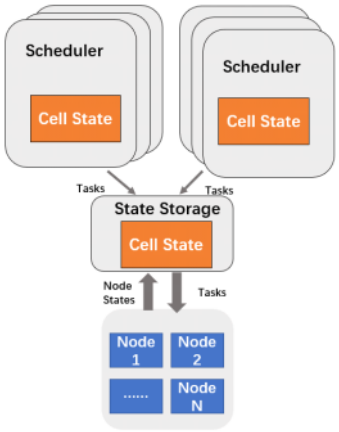
- State Storage 模块（资源维护模块）负责存储和维护资源及任务状态，以便 Scheduler 查询资源状态和调度任务；
- Resource Pool 即为多个节点集群，接收并执行 Scheduler 调度的任务；
- 而 Scheduler 只包含任务调度操作，而不是像单体调度器那样还需要管理集群资源等。

与两层调度的不同.

- 存在多个调度器，每个调度器都可以拥有集群全局的资源状态信息，可以根据该信息进行任务调度；
- 共享状态调度是乐观并发调度，在执行了任务匹配算法后，调度器将其调度结果提交给 State Storage，由其决定是否进行本次调度，从而解决竞争同一种资源而引起的冲突问题，实现全局最优调度。而，两层调度是悲观并发调度，在执行任务之前避免冲突，无法实现全局最优匹配。

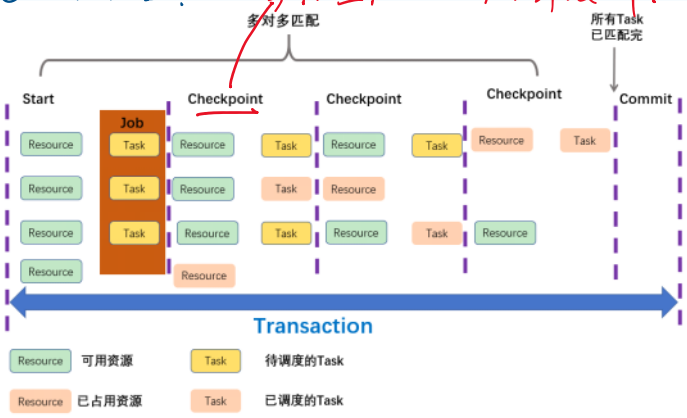
设计. (Omega).

① 架构设计.



② 工作原理.

检查 Resource 是否都被占用.



这里的 Job 相当于一个事务，也就是说，当所有 Task 匹配成功后，这个事务就会被成功 Commit，如果存在 Task 匹配不到可用资源，那么这个事务需要执行回滚操作，Job 调度失败。

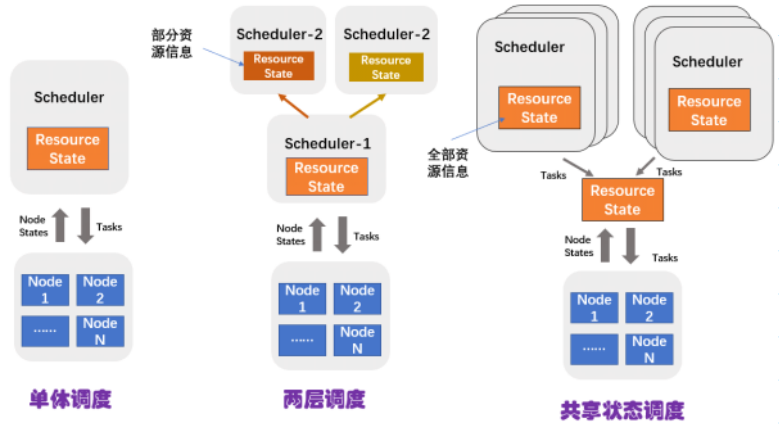
→ 核心问题.

Commit，如果存在 Task 匹配不到可用资源，那么这个事务需要执行回滚操作，Job 调度失败。

核心问题.

③ Job并发调度. 借用DB中的MVCC思想.

三种调度方式对比.



	单体调度	两层调度	共享状态调度
调度架构	集中式结构：一个中央调度器	树形结构：一个中央调度器，多个第二层调度器	分布式结构：多个对等调度器
调度形式	单点集中调度	Resource Offer	Transaction
调度单位	Task	Task	Task
任务调度的并发性	无并发	悲观并发调度	乐观并发调度
是否是全局最优调度	是	否	是
系统并发度	共享状态调度>两层调度>单体调度		
调度效率（综合考虑并发度，全局最优性，以及故障问题等因素）	共享状态调度>两层调度>单体调度		
系统可扩展性	共享状态调度>两层调度>单体调度		
是否有具体实现源码	是	是	否
适用场景	小规模集群，适用于业务类型比较单一的场景	中等规模集群，适用于同时具有多种业务类型的场景	大规模集群，适用于同时具有多种业务类型的场景
典型应用	Borg	Mesos、YARN	Omega